基于广义分形插值理论的多尺度分类尺度下推算法 *

李佳星 [a, b, c]，赵书良 [a, b, c]，安　磊 [a, b, c]，李长镜 [a, b, c]

(河北师范大学 a. 数学与信息科学学院; b. 河北省计算数学与应用重点实验室; c. 移动物联网研究院，石家庄 050024)

摘　要：多尺度数据挖掘多应用于空间遥感图像数据，以图像的分辨率或者区域分割为依据进行尺度划分，然后在每个尺度层进行分析。近期，有不少学者将多尺度数据挖掘应用于一般数据集上，以等级理论、概念分层以及包含度理论等为尺度划分依据，研究不同尺度层的分布规律，进而发现有意义的事实，如多尺度关联规则以及多尺度聚类。但是在一般数据集下，很少将多尺度数据挖掘应用于分类算法领域。定义了广义分形插值理论的概念，打破了局限于迭代函数系统 IFS（iterative function systems）的缺憾，拓展了分形插值的应用；提出了基于广义分形插值理论的多尺度分类尺度下推算法 MSCSDA（multi-scale classification scaling-down algorithm）。仿真实验建立在四个 UCI 基准数据集和一个 H 省部分人口真实数据集上，并将 MSCSDA 与 KNN、Decision Tree 以及 LibSVM 算法进行对比分析，实验结果表明，MSCSDA 算法在不同的数据集上均优于其他算法。

关键词：多尺度数据挖掘；分类；分形插值；尺度下推

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.01.0031

Scaling-down algorithm of multi-scale classification based on generallized fractal interpolation theory

Li Jiaxing [a, b, c], Zhao Shuliang [a, b, c], An Lei [a, b, c], Li Changjing [a, b, c]

(*a. Mathematics & Information Science College*, *b. Hebei Key Laboratory of Computational Mathematics & Applications*, *c. Institute of Mobile Internet of Things Hebei Normal University*, *Shijiazhuang* 050024, *China*)

Abstract: The research of multi-scale data mining mainly applied to space remote sensing image data sets, and conduct scale division based on the resolution or regional segmentation of the images, then analysis knowledge on each scale layer. Recently, there are quite a few learners applied the multi-scale data mining to general data sets, and conduct scale division based on the level theory, concept hierarchy and inclusion degree etc. , study the distribution rule on different scale layers, and then found significant facts. For example, multi-scale association rules, multi-scale clustering. But it has not been involved in the field of the classification mining. This paper defines the concept of generalized fractal interpolation theory, break the situation that limited to the use of the iteration function system(IFS) , and extend the application of the fractal interpolation. Then, a multi-scale classification scaling-down algorithm based on the generalized fractal interpolation theory named MSCSDA (Multi-Scale Classification Scaling-Down Algorithm) is proposed. This paper performs experiments on four UCI benchmark data sets, and one real data set (H province part of the population) . Then analysis the experimental results compare MSCSDA with KNN, Decision Tree and LIBSVM algorithms on different data sets. The experimental results show that the MSCSDA algorithm gives better results in terms of classification than the others.

Key words: multi-scale data mining; classification; fractal interpolation; scale-down

0　引言

尺度的定义来源于地学科学，一般指在学习分析中所涉及的空间或时间单位，也可以指在空间或时间上，某一个过程或现象所发生的范围或频率。研究表明，客观世界中普遍存在尺度现象[1]。

多尺度数据挖掘多应用于空间遥感图像数据，以图像的分辨率或者区域分割为依据进行尺度划分，然后在每个尺度层进

收稿日期：2018-01-18；**修回日期：**2018-03-06　　**基金项目：**国家自然科学基金资助项目（71271067）；国家社科基金重大项目（13&ZD091）；河北省高等学校科学技术研究项目（QN2014196）；河北师范大学硕士基金资助项目（xj2015003）

作者简介：；李佳星，（1992-），女，河北石家庄人，硕士研究生，主要研究方向为数据挖掘、智能信息处理；赵书良（1967-），男（通信作者），河北献县人，教授，博导，主要研究方向为数据挖掘、智能信息处理（zhaoshuliang@sina.com）；安磊，（1991-），男，硕士研究生，主要研究方向为数据挖掘、智能信息处理.；李长镜，（1990-），男，硕士研究生，主要研究方向为数据挖掘、智能信息处理.

行分析。近期，有不少学者将多尺度数据挖掘应用于一般数据集上，以等级理论、概念分层以及包含度理论等为尺度划分依据，研究不同尺度层的分布规律，进而发现有意义的事实，如多尺度关联规则以及多尺度聚类[2]。

但是在一般数据集下，很少将多尺度数据挖掘应用于分类算法领域。目前有学者提出基于分形理论的多尺度分类挖掘研究方法，其详细论述了多尺度数据挖掘与分类算法相结合的有效性与可行性，介绍了多尺度分类的基本概念以及基本任务；并从多方面阐述了多尺度数据集具有分形特性，即自相似性、标度不变性、自仿射性以及层次性。进而将衡量分形自相似结构的指标即分形维数作为尺度转换算法的方法论，提出基于分形理论的多尺度分类尺度上推算法；尺度转换算法作为多尺度分类数据挖掘的核心研究内容，本质上是一个知识推演的过程，即由一个指定尺度上学习得到的知识或者信息推演归算到另一个尺度上，包括两部分即尺度上推算法和尺度下推算法，两个算法的使用由实际应用中目标尺度相对于基准尺度的定性大小决定的。因而两个算法相应而生，相辅相成。在传统的遥感图像数据下，对尺度下推算法的研究已经具备了较为成熟的理论和方法，然而在一般数据集下，尺度下推算法作为尺度转换算法中不可或缺的一部分，目前相关研究较少，理论与方法均有待完善。

尺度下推就是根据大尺度上得到的知识，结合小尺度固有的信息，推演出小尺度上的知识。本质上讲，是一个由模糊到精确、弱化整体特征，加强局部特征、忽略宏观特征，保留微观特征以及信息由少变多的过程。尺度下推的关键在于如何增加小尺度细节信息，插值是最常用的方法。

最近邻插值法也称为零阶插值，以距离待估样本最近的样本值作为插入值，是最简单的插值方法，但是精确度不高，尤其当数据的细微差距较大时。反距离权重插值的本质在于以多个已知样本的线性组合作为待估样本值，在尺度下推的应用中，将大尺度上学习到的知识作为已知样本，但是并未考虑小尺度上的数据所固有的分布细节特征。样条插值法加强了紧邻的两个数据点间的细节，可以学习到分布光滑的模型，但是在处理复杂的数据时，可能会遗漏局部细节结构的信息。克里格插值法以及双线性插值法都是应用十分广泛的插值法，都是基于待插点四周的已知样本点的信息，但并未考虑整体分布的趋势所决定的待估样本应具有的独特差异[3,4,5]。分形插值根据自相似这一特性，既考虑了整体的分布趋势，又加强了局部分布独有的特征，为处理非线性且分布复杂的数据提供了新的思路[6]。

本文借助分形插值的理论，定义了广义分形插值理论的概念，打破了局限于迭代函数系统 IFS（iterative function systems）的缺憾，拓展了分形插值的应用；进而提出了基于广义分形插值理论的多尺度分类尺度下推算法 MSCSDA。

1　分形插值

分形插值为处理大量分布离散且不光滑的数据提供了新思路，是解决非线性数据的有效工具。传统分形插值的核心在于根据自相似性，构造迭代函数系统 IFS[7]，由已知的样本迭代出新样本。目前多应用于具有自相似性结构的事物的建模、仿真以及数据可视化。但是仅局限于三维以下的数据，大大限制了分形插值在一般多维数据下的应用。

从广义上讲，就是根据自相似性，从不同尺度层面衡量已知样本对待估样本的贡献，既要考虑整体分布的趋势，又要考虑待估样本附近的局部已知样本的分布特点。

定义 1　广义分形插值尺度划分。设 $f:(X)\to Y$，其中 X 为已知样本点，经过函数 f 映射，得到待估样本 x_0 的候选样本集 Y。函数 f 用一个三元组表示 (L,X,W)。其中，$L=(l_1,l_2,\cdots,l_n)$，表示根据某种条件定义 n 个尺度层；$X=(x_1,x_2,\cdots,x_n)$，表示 n 个尺度层下包含的已知样本的子集；$W=(w_1,w_2,\cdots,w_n)$，表示 n 个尺度层下包含的已知样本的子集对待估样本 x_0 的贡献，其中 $w_i=\varphi(x_i)$；在每个尺度层下，由已知样本的子集，根据其对待估样本的贡献，得到一个待估样本 x_0 的候选样本。

定义 2　广义分形插值。设 $\Gamma:(Y)\to x_0$，其中，待估样本 x_0 的候选样本集 Y，经过机制 Γ 得到最终的待估样本。

2　多尺度分类

多尺度分类作为一个跨学科研究课题，其实质是将多尺度科学与分类相结合，多尺度、全方面地研究数据特征，从而得到不同层面的分类模型，进而研究尺度转换机制以及尺度转换引起的尺度效应问题。构造多尺度数据集是进行多尺度数据挖掘的第一步，进而在不同尺度层上对数据进行多角度的分析学习。尺度转换作为研究的核心，其实质是由某一尺度层数据集上学习得到的知识推演得到其他尺度层上的知识，目的在于一次学习多次利用，避免繁琐的学习过程。

2.1　构造多尺度数据集

构造多尺度数据集是一个由整变零的过程。将原始数据集划分为具有一定偏序关系的多个子数据集，且同层数据集互不相交。

包含度理论以及等级理论一般用于处理模糊、不确定性关系，将连续的问题离散化为具有一定偏序关系的问题，逐层深化解决[8]。这也正是构造多尺度数据集的思路。因此，本文引入包含度理论以及等级理论，将数据集的某一个或多个特征属性值离散化为具有偏序关系的范围，由此构造多尺度数据集。如图 1 所示，是一个四层多尺度数据集，其结构类似于树，根节点代表原始数据集，节点中的 f 标志，代表在该数据集下训练得到的分类模型。

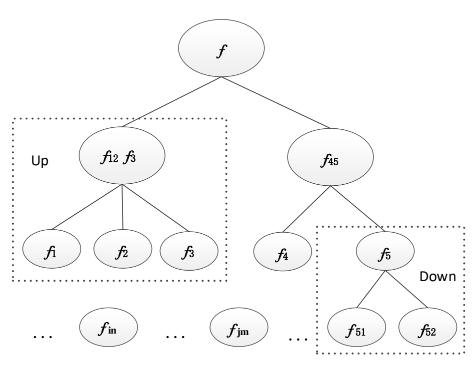


图 1 四层多尺度数据集

本文所提到的偏序关系是绝对的偏序关系，即图 1 中每个节点仅与它的孩子节点和双亲节点有关系，兄弟节点之间关系并不紧密。如图 1 所示，选择第三尺度层为基准尺度，若选择第二尺度层为目标尺度，则尺度上推发生在第二尺度层的每一个节点以及与它有关系的第三尺度层的节点上，如图 1 Up 部分，是一个由多变少的过程；若选择第四尺度层为目标尺度，则尺度下推发生在第三尺度层的每一个节点以及与它有关系的第四尺度层的节点上，如图 1Down 部分，是一个由一变多的过程。

2.2 多尺度分类

多尺度数据挖掘的核心在于尺度转换，即由基准尺度层得到的知识推演出目标尺度层上的知识。通过以上分析，定义在一般数据集下多尺度分类的概念：

定义 3 多尺度分类。由原始数据集，根据包含度理论以及等级理论，构造出多尺度数据集；由某些评价指标确定基准尺度层数据集；采用传统分类方法在基准尺度层数据集上来训练分类模型；确定尺度转换机制，由基准尺度层数据集上训练的分类模型推演出目标尺度层数据集的分类模型。

3 多尺度分类尺度下推算法 MSCSDA

对于多尺度分类而言，每层尺度上学习到的知识是分类模型。那么尺度下推就是由整体数据集上训练的分类模型推演出划分后的各个局部数据集上的分类模型，是一个由少变多的过程。

本文以 SVM 一对一多类别分类算法作为基准分类方法，那么，基准尺度层上学习得到的知识即分类模型，包含的信息有支持向量、权重系数以及常数 b。本文将大尺度上学习到的分类模型作为小尺度上的基本分类模型，其中的权重信息以及常数 b 保持不变，不同的是支持向量。因此本文将支持向量作为尺度转换的对象，根据包含度理论、等级理论进行尺度划分后，大尺度上学习得到的支持向量，会被划分到小尺度上的各个局部划分中去，本文将某一局部划分中保留的一部分大尺度上学习到的支持向量作为已知样本，而缺失的另一部分作为待估样本，小尺度上的划分后的数据信息以及紧邻待估样本的已知样本作为局部细节结构。已知样本也就是模糊的分类边界信息，代表中宏观的整体趋势，局部细节结构代表着微观细节信息，那么多尺度分类尺度下推的目的就是使得模糊的分类边界精确化。

3.1 MSCSDA 实现思想

多尺度分类尺度下推的思想如下。首先，构建多尺度数据集，本文依据包含度理论与等级理论，构造出具有树结构的多尺度数据集，每一个局部划分节点都仅和它的双亲、孩子节点有直接关系。其次，选择基准尺度，并在基准尺度层采用一对一 SVM 分类算法训练得到基准分类模型，记录有效知识，权重 W 以及常数 b,还有作为尺度转换对象的支持向量 X。然后，根据多尺度分类尺度下推实现机制推演出目标尺度层每个局部划分的分类模型。

3.2 MSCSDA 理论基础

3.2.1 广义分形插值理论

根据第 1 章定义的广义分型差值理论，从不同尺度层面衡量已知样本对待估样本的贡献，既要考虑整体分布的趋势，又要考虑待估样本附近的局部已知样本的分布特点。

3.2.2 反距离加权[9]

根据已知样本与待估样本间的距离衡量已知样本对估计待估样本的贡献，距离越近，贡献越大，反之越小。

最常用的权重定义公式：

$$w_i = \frac{d_{oi}^{\ -p}}{\sum_{j=1}^{n} d_{oj}^{\ -p}} \tag{1}$$

其中，d_{oi} 表示待估样本与已知样本的距离；p 为任意正实数，一般取 p=1 或 p=2。

3.3 MSCSDA 实现机制

根据以上分析可以看出，对于目标尺度层的每一个局部划分会产生多个待估样本，对于每个待估样本 $x_o^{'}$ 都需要进估计，其中 x_o 代表局部划分中缺失的已知样本。

详细步骤如下：

a)筛选出与待估样本 $x_o^{'}$ 类别一致的已知样本集 A 和局部划分中除已知样本外的其他样本集 B。支持向量从本质上讲是类别边界样本，所以要选类别一致的数据进行分析插值。

b)根据广义分形插值尺度划分，结合已知样本 A 的数据值特点，确定 Nc 个尺度，求出每个尺度下覆盖的已知样本集 A_i，每个尺度下都会求得一个待估样本的候选样本 y_i；然后根据广义分形插值 $\Gamma:(Y)\to x_0$，求得最终待估样本值，其中机制 Γ 表示如下：

$$x_o^{'} = \frac{1}{N_c}\sum_{i=1}^{N_c} y_i \tag{2}$$

c)求待估样本 $x_o^{'}$ 的候选样本 y

其中，Ai 记作 $C=(c_1,c_2,\cdots,c_n)$，包含 n 个已知样本。

(a)为 C 的每个已知样本加权重值。根据理论基础的反距离加权法得到权重的最终形式，这里 p 取值为 1。

$$w_i=\frac{d_{oi}^{\ -1}}{\sum_{j=1}^{n} d_{oj}^{\ -1}} \qquad (3)$$

其中 d_{oi} 为待估样本与已知样本的欧式距离。

(b)求候选样本 x。y 即为局部划分中除已知样本外的其他样本集 B 中，与已知样本 C 相似性最高的样本。

$$y=\arg\max_{b\subseteq B} S(C,b) \qquad (4)$$

其中相似性度量方法

$$S(C,b)=\sum_{i=1}^{n}\frac{w_i}{\|c_i-b\|} \qquad (5)$$

与具有较大权重的已知样本越近的样本，越有可能成为候选样本。

3.4 MSCSDA 伪代码

Algorithm： MSCSDA 算法

Input： Original Datasets

Output： The Classification Model of Target Scale

Data Preprocessing

Data Scaling; //尺度划分；

Build multi-scale datasets; //构建多尺度数据集；

Get knowledge on the basic scale

Choose the basic scale of $BS(d_{BS}^1,d_{BS}^2,\ldots,d_{BS}^m)$；

//选择基准尺度；

foreach d_{BS}^i do begin

Classifing on sub datasets;

Get weight matrix;

Get constant b;

Get support vector A;

end for

Scale transformation

foreach sub dataset B do begin

Get Nc scale layers;

//根据分形插值理论确定 Nc 个尺度层；

foreach scale layer do begin

Get the known sample set C; //确定已知样本；

Get the weight of each known samp based the inverse distance weighting formula;

$$w_i=\frac{d_{oi}^{\ -1}}{\sum_{j=1}^{n} d_{oj}^{\ -1}}$$ //为已知样本赋权重

Get similarity of each sample in local division

$$S(C,b)=\sum_{i=1}^{n}\frac{w_i}{\|c_i-b\|}$$

//衡量局部划分中样本与已知样本的相似性

Get candidate of the estimated sample y;

$$y=\arg\max_{b\subseteq B} S(C,b)$$ //得到候选样本 y

end for

Get final value of the estimated sample $x_o^{'}$

$$x_o^{'}=\frac{1}{N_c}\sum_{i=1}^{N_c} y_i$$ //确定最终的待估样本值

end for

return The Classification Model of Target Scale

4 实验分析

SVM 算法是解决分类问题最常用的算法，针对多类别分类问题，常用的算法有一对一和一对多形式。本文采用的基本分类算法是 SVM 一对一形式多类别分类算法，其中核函数采用 RBF。本文使用 MATLAB 实现 MSCSDA 算法，借助 LIBSVM 库，其默认的便是一对一形式，因此本文后续称 SVM 一对一多类别分类算法为 LIBSVM 算法。MSCSDA 算法是建立在 LIBSVM 算法之上的。

本文在四个 UCI 公共数据集以及一个真实数据集上进行实验，并与 Decision Tree、KNN 以及 LIBSVM 算法进行实验对比，验证本文 MSCSDA 算法的有效性与可行性。

4.1 数据集

本文采用的四个 UCI 公共数据集包括 Ionosphere 数据集、Pima Indians Diabetes（PID）数据集、Spambase 数据集以及 wine 数据集，真实数据集采用 H 省部分人口数据，这五个数据集的样本数量、特征属性以及类别标签数量都不尽相同。详细信息见表 1 数据集的详细信息。

表 1 数据集的详细信息

数据集	样本数	特征数	类别数
Ionosphere	351	34	2
PID	768	8	2
Spambase	4601	57	2
wine	178	13	3
H 省部分人口数据	6311	7	3

4.2 评价指标

本文实验中，采用最常用的四个评价指标（正确率

（Accuracy）[10]、标准化互信息（NMI）[11]、F1-Measure 以及运行时间（Run Time））来衡量 MSCSDA 算法的分类性能以及体现多尺度分类的优势。

4.2.1 Accuracy（Acc）

分类的正确率表示两者之间的一一对应关系，即正确对应的样本个数占全部样本的比例，计算公式如下：

$$Acc = \frac{1}{n}\sum_{i=1}^{n}\delta(C_i, map(P_i)) \quad (6)$$

其中：n 为样本总数量，c_i 为第 i 个样本数据的真实类别标签，$map(P_i)$ 表示第 i 个样本的实验结果 p_i 到真实类别标签的最优映射，$\delta(x,y)$ 是一个匹配函数，当 x=y 时，$\delta(x,y)=1$，否则 $\delta(x,y)=0$。Acc 值越高，表示分类效果越好。

4.2.2 NMI

标准化互信息（NMI）一般借助混淆矩阵计算求得，计算公式如下：

$$NMI = \frac{\sum_{ij}\frac{n_{ij}}{n}\cdot\log\frac{n\cdot n_{ij}}{n_i\cdot n_j}}{\sqrt{(\sum_i n_i\cdot\log\frac{n_i}{n})(\sum_j n_j\cdot\log\frac{n_j}{n})}} \quad (7)$$

其中：n_i 为真实标签为 i 的样本数量，n_j 为在实验中预测的标签为 j 的样本数量，n_{ij} 标识号真实标签为 i，但在实验中预测的标签为 j 的样本数量。NMI 的值越高，表示分类效果越好。NMI 与 Acc 有一定的关系，当 Acc 值缓慢下降时，NMI 值会迅速下降，当两个 Acc 值很近似时，其 NMI 值会更近似，因此，NMI 不仅放大了 Acc 差异，还增强了 Acc 的相似度。

4.2.3 F1-measure

F_1-measure 作为一个重要评价指标，其值越大，说明分类性能越好。对于多类分类问题，F1-Measure 计算公式如下：

$$F_1\text{-measure} = \frac{1}{c}\sum_{i=1}^{c}F_1\text{-measure}_i \quad (8)$$

其中 $F_1\text{-measure}_i$ 采用一对多的方法，即将第 i 类样本作为正类，其余类样本作为负类，产生的 c 个 F_1-measure 值，求和取均值。

4.3 实验结果分析

本文首先根据包含度理论以及等级理论，将数据集的一个特征值离散化为不同的范围，将数据集划分为二层尺度的多尺度数据集，第一层为原始数据集，第二层划分为二到五部分不等，如 H 省部分人口数据，是按照区域代码划分为两部分数据集，如图 1 所示 Down 部分。

本文首先在不同尺度层数据集上比较分类算法（KNN、decision tree 和 LIBSVM）的各评价指标，体现多尺度分类的优势；其次对比分析本文提出的 MSCSDA 算法的性能。

表 2 各分类算法的 Acc 值结果 / %

数据集	KNN		decision tree		LIBSVM		MSCSDA
	第一层	第二层	第一层	第二层	第一层	第二层	
Ionosphere	73.50	75.21	75.21	76.92	74.36	75.21	**77.78**
PID	72.90	74.07	71.35	73.10	74.27	74.46	**75.05**
Spambase	81.83	82.43	79.35	81.00	77.04	79.96	**84.04**
wine	90.91	93.18	84.09	86.36	93.18	93.18	**95.45**
H 省部分人口数据	92.42	93.99	94.01	96.39	94.82	96.93	**98.08**

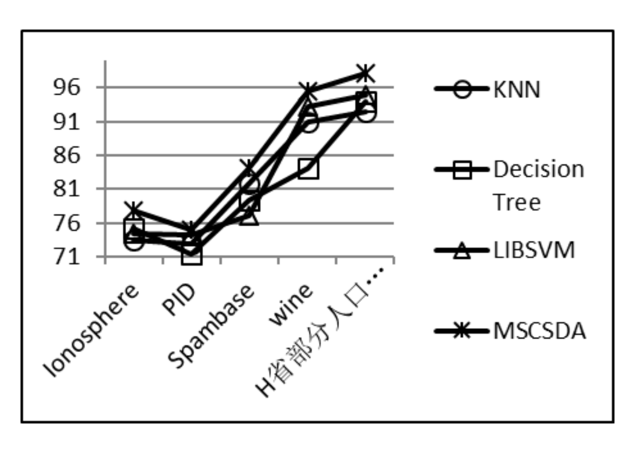


图 2 各分类算法在第一层数据集上的 Acc 值

表 2 展示的是各分类算法在两个尺度层数据集上的 Acc 值。从表 2 的结果中可以明显看出，各分类算法在第二尺度层数据集上的 Acc 值较第一层有显著提升，平均提升大约 2%。

主要原因可能在于，整个数据集的分布呈现为无规律状态，训练的分类模型复杂多样，即使分类的 Acc 值很高，也极有可能出现过拟合的问题。但是数据集经过多尺度的划分后，降低了各子数据集的分布复杂性，同时也就降低了学习到的分类模型的复杂性。而且本文提出的多尺度划分是建立在经验的包含度和等级理论之上的，目标性更强。文献[12]提出了基于聚类的 Boosting 方法（CBB），其主要思想是先聚类再分类，但是应用的数据集仅局限于球形分布的数据集，而本文提出的多尺度分类思想对数据集的类型不做限制。

本文提出的 MSCSDA 方法，以 LIBSVM 算法为基准分类算法，选择 LIBSVM 算法下第一层即原始数据集为基准尺度数据集，经过 MSCSDA 算法，得到下推后的第二尺度层的分类模型。从表 2 中看到，MSCSDA 算法的 Acc 值较 LIBSVM 的第一层具有显著的提升，平均大约提升了 3%，并且经过 MSCSDA 算法得到的第二尺度层的 Acc 与在第二尺度层数据集上直接训练的分类模型的 Acc 值相比较，平均大约提升了 2%。

图 3 呈现的是各分类算法在第一尺度层数据集上的 Acc

值对比，从图中也可以明显看出 MSCSDA 算法较其他单一分类算法有明显的优势；

表 3 各分类算法的 NMI 值结果

数据集	KNN		Decision Tree		LIBSVM		MSCSDA
	第一层	第二层	第一层	第二层	第一层	第二层	
Ionosphere	0.1900	0.2690	**0.2690**	**0.1820**	0.2532	0.2690	**0.3172**
PID	0.1053	0.1237	0.1040	0.1130	0.1301	0.1303	**0.1394**
Spambase	0.3071	0.3130	0.2692	0.2959	0.2261	0.2941	**0.3874**
wine	0.7384	0.7854	0.5898	0.6229	0.7897	0.7897	**0.8709**
H 省部分人口数据	0.7574	0.7893	0.8160	0.8806	0.7828	0.8804	**0.9014**

图 3 各分类算法在第一层数据集上的 NMI 值

标准互信息 NMI 与 Acc 有一定的关系，当 Acc 值缓慢下降时，NMI 值会迅速下降，当两个 Acc 值很近似时，其 NMI 值会更近似，因此，NMI 不仅放大了 Acc 差异，还增强了 Acc 的相似度。从表 3 中可以看出，经过多尺度划分后，各分类算法在第二尺度层数据集上的 NMI 值较第一层有显著提升，平均提升大约 0.03。而 Libsvm 算法在 H 省部分人口数据的第二层数据集上较第一层数据集上的 NMI 值大约提升了 0.1。MSCSDA 算法的 NMI 值较 LIBSVM 算法在第一层数据集上的 NMI 均有显著提升，尤其是在 H 省部分人口数据数据集上，提高了大约 0.1。

图 3 展示的是各分类算法在第一层数据集上的 NMI 值，从中可以看出， MSCSDA 算法的 NMI 值均高于其他算法。

值得一提的是在 Ionosphere 数据集上，Decision Tree 算法，第二层的 NMI 值较第一层略有下降，可以看出看不同尺度下详细的混淆矩阵分布情况。

表 4　不同尺度层下的混淆矩阵

第一层混淆矩阵		第二层混淆矩阵	
73	0	64	9
29	15	18	26

从表 4 中可以看出，第一层很明显将全部的第一类数据正确分类，而有 29 条错误均是将第二类数据误判为第一类；而第二层很明显的两类均有错分数据，但是总的错分数量较少些，也就是 Acc 值较高。根据 NMI 的计算公式可知，越是能将更多的类别数据全部正确分类的模型，NMI 值越高。但是很明显，在 Ionosphere 这个数据集上，分类边界是个模糊复杂边界，第一层将模糊边界的数据全部划分给第一类，而第二层则是取的中间界，尽管第一层的 NMI 值高一些，但是以一概全的方式并不可取。

表 5 各分类算法的 F_1 值结果

数据集	KNN		decision tree		LIBSVM		MSCSDA
	第一层	第二层	第一层	第二层	第一层	第二层	
Ionosphere	0.6487	0.6714	0.6714	**0.7420**	0.6562	0.6714	0.7147
PID	0.6515	0.6873	0.6843	0.6866	0.6944	0.6973	**0.7003**
Spambase	0.8014	0.8132	0.7662	0.7892	0.7359	0.7706	**0.8215**
wine	0.9095	0.9343	0.8390	0.8629	0.9299	0.9299	**0.9508**
H 省部分人口数据	0.9103	0.9229	0.9310	0.9553	0.9345	0.9579	**0.9714**

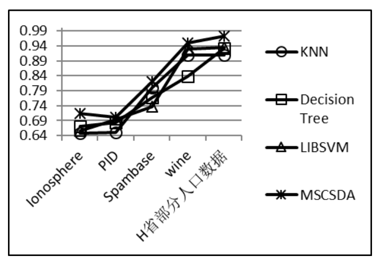


图 4 各分类算法在第一层数据集上 F1 值

F_1-measure 作为一个重要评价指标，其值越大，说明分类性能越好。如表 5 所示，经过多尺度划分后，各分类算法在第二尺度层数据集上的 F_1 值较第一层有显著提升，并且第二层较第一层的 F1 平均提升大约 0.02，尤其是 Ionosphere 数据下 decision tree 算法，提升了约 0.07；而本文提出的 MSCSDA 算法，在多数数据集下，其 F_1-measure 也是最高的，虽然在第二层 Ionosphere 数据集下，仅次于 decision tree 算法，但是相

对于 LIBSVM 第一层数据集下的 F1-Measure 值，已经提高了平均约 0.06。如图 4 所示，是各分类算法在第一层数据集上的 F1-Measure 值，本文提出的 MSCSDA 算法的 F_1-measure 值均高于其他算法。

表 6 各分类算法在第二层数据集上的运行时间（run time）/s

数据集	KNN	decision tree	LIBSVM	MSCSDA
wine	0.007	0.018	0.004	**0.002**
Ionosphere	0.009	0.027	0.008	**0.005**
PID	0.011	0.031	0.006	**0.004**
Spambase	0.079	0.149	0.2340	**0.003**
H 省部分人口数据	0.111	0.028	0.118	**0.008**

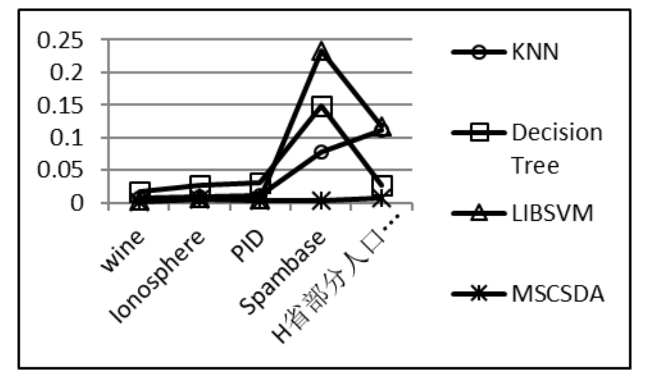


图 6 各分类算法的运行时间（run time）/s

运行时间是评价一个算法是否高效可行的一个关键指标。从表 6 与图 6 中可以明显看出，KNN、decision tree 以及 LIBSVM 算法的 run time 值随着各数据集的样本数量、特征属性以及类别标签的增加呈现出上升趋势，特别是在 Spambase 数据集上，LIBSVM 算法和 decision tree 算法的 run time 发生了较大的波动。而本文提出的 MSCSDA 算法，由于省去了训练学习过程，仅需要由基准尺度层得到的分类模型经过上推机制得到，因此 run time 值始终处于较低的水平，且波动不大。

综合而言，通过对以上实验结果的对比分析，本文提出的多尺度分类思想以及尺度下推 MSCSDA 算法相对于其他单一的分类算法具有显著的优势，同时验证了该算法的有效性与可行性。

5 结束语

本文针对目前基于分形理论的多尺度分类挖掘研究中尚未解决的尺度下推算法，借助分形插值的理论，定义了广义分形插值理论的概念，打破了局限于迭代函数系统 IFS 的缺憾，拓展了分形插值的应用；进而提出了以 LIBSVM 为基准分类模型的基于广义分形插值理论的多尺度分类尺度下推算法 MSCSDA。以不同粗细程度的尺度衡量已知样本对未知样本的估计所作出的贡献，既考虑了已知样本整体分布的趋势，又考虑了待估样本附近的局部已知样本的分布特点，使得模糊的分类边界更精确。

在接下来的研究工作中，将致力于研究多尺度分类挖掘的理论支撑，拓展其他分类方法（决策树、贝叶斯、神经网络等）的转换对象，寻求更优的尺度转换机制，衡量基准尺度选择的评价指标，从而完善多尺度分类挖掘的理论和方法。

参考文献：

[1] 韩玉辉，赵书良，柳萌萌，等．多尺度聚类挖掘算法 [J]. 计算机科学, 2016, 43 (8): 244-248.

[2] 苏东海，赵书良，柳萌萌，等．基于加权向量提升的多尺度聚类挖掘算法 [J]. 计算机科学, 2015, 42 (4): 263-267.

[3] 李正泉，吴尧祥．顾及方向遮蔽性的反距离权重插值法 [J]. 测绘学报, 2015, 44 (1): 91-98.

[4] 梁永忠，葛咏，王江浩．基于地统计学的尺度下推方法综述 [J]. 遥感技术与应用, 2015, 30 (1): 1-7.

[5] Wang Qunming, Shi Wenzhong, Atkinson P M, *et al.* A new geostatistical solution to remote sensing image downscaling [J]. IEEE Trans on Geoscience and Remote Sensing, 2015, 1-11.

[6] Liu Liqiang, Wang Xiangguo, Ren Huili. 3D seabed terrain establishment based on moving fractal interpolation [C]// Proc of the 7th International Joint Conference on Computational Sciences and Optimization. Washington DC: IEEE Computer Society, 2014: 6-10

[7] 刘红艳，陈宇坤，李信富．地震道重建和重采样的分形插值方法研究 [J]. 地球物理学进展, 2014, 29 (2): 518-522.

[8] 赵泉华，刘冬，李晓丽，等．利用包含度和隶属度的遥感影像模糊分割 [J]. 中国图象图形学报, 2017, 22 (7): 988-995.

[9] 李晓晖，袁峰，贾蔡，等．基于反距离加权和克里格插值的 S-A 多重分形滤波对比研究 [J]. 测绘科学, 2012, 37 (3): 87-89+46.

[10] Chen Wenyen, Song Yangqiu, Bai Hongjie, *et al.* Parallel spectral clustering in distributed systems [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2011, 33 (3): 568-586.

[11] Allab K, Labiod L, Nadif M. A semi-NMF-PCA unified framework for data clustering [J]. IEEE Trans on Knowledge and Data Engineering, 2017, 29 (1): 2-16.

[12] Dee Miller, Leen-Kiat Soh. Cluster-Based Boosting [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27 (6): 1491-1504.